

(19)日本国特許庁(JP)

(12)公開特許公報(A)

(11)特許出願公開番号

特開平5-346797

(43)公開日 平成5年(1993)12月27日

(51)Int.Cl. ⁵	識別記号	庁内整理番号	F I	技術表示箇所
G 1 0 L 9/00		C 8946-5H		
G 0 6 F 15/36		D 8320-5L		
G 1 0 L 3/00	5 1 3	D 8842-5H		
	5 1 5	D 8842-5H		

審査請求 未請求 請求項の数8(全 20 頁)

(21)出願番号 特願平5-828

(22)出願日 平成5年(1993)1月6日

(31)優先権主張番号 特願平4-121460

(32)優先日 平4(1992)4月15日

(33)優先権主張国 日本(JP)

(71)出願人 000002185

ソニー株式会社

東京都品川区北品川6丁目7番35号

(72)発明者 西口 正之

東京都品川区北品川6丁目7番35号 ソニー株式会社内

(72)発明者 松本 淳

東京都品川区北品川6丁目7番35号 ソニー株式会社内

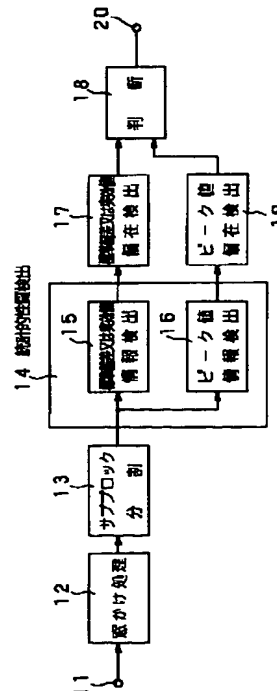
(74)代理人 弁理士 小池 晃 (外2名)

(54)【発明の名称】 有声音判別方法

(57)【要約】

【構成】 入力信号の1ブロック(1フレーム)をサブブロック分割部13でさらに分割し、統計的性質検出部14でサブブロック毎に標準偏差又は実効値の情報とピーク値情報とを検出する。偏在検出部17では標準偏差又は実効値の時間軸上での偏在を検出し、偏在検出部19ではピーク値の偏在を検出して、これらの偏在検出情報に基づいて判断部18が各ブロック毎に有声音か無声音かを判別する。

【効果】 有声音か無声音(又はノイズ)かの判別が確実に行える。



【特許請求の範囲】

【請求項1】 入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、

1ブロックの信号を複数のサブブロックに分割する工程と、

上記複数のサブブロック毎に信号の統計的な性質を求める工程と、

上記統計的な性質の時間軸上での偏りに応じて有声音か否かを判別する工程とを有することを特徴とする有声音判別方法。

【請求項2】 上記信号の統計的な性質とは各サブブロック毎の信号のピーク値、実効値又は標準偏差であることを特徴とする請求項1記載の有声音判別方法。

【請求項3】 入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、

1ブロックの信号の周波数軸上のエネルギー分布を求める工程と、

上記1ブロックの信号のレベルを求める工程と、

上記1ブロックの信号の周波数軸上のエネルギー分布と信号レベルとに応じて有声音か否かを判別する工程とを有することを特徴とする有声音判別方法。

【請求項4】 入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、

1ブロックの信号を複数のサブブロックに分割する工程と、

上記複数のサブブロック毎の信号のピーク値、実効値又は標準偏差の時間軸上での偏りを求める工程と、

1ブロックの信号の周波数軸上のエネルギー分布を求める工程と、

上記1ブロックの信号のレベルを求める工程と、

上記複数のサブブロック毎の信号のピーク値、実効値又は標準偏差の時間軸上での偏りと上記1ブロックの信号の周波数軸上のエネルギー分布又は上記1ブロックの信号のレベルとに応じて有声音か否かを判別する工程とを有することを特徴とする有声音判別方法。

【請求項5】 入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、

1ブロックの信号を複数のサブブロックに分割する工程と、

上記複数のサブブロック毎に時間軸上で信号のピーク値、実効値又は標準偏差を求める工程と、

上記1ブロックの信号の周波数軸上のエネルギー分布を求める工程と、

上記1ブロックの信号のレベルを求める工程と、

上記複数のサブブロック毎の信号のピーク値、実効値又は標準偏差と上記1ブロックの信号の周波数軸上のエネ

ルギー分布と上記1ブロックの信号のレベルとに応じて有声音か否かを判別する工程とを有することを特徴とする有声音判別方法。

【請求項6】 入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、

1ブロックの信号を複数のサブブロックに分割する工程と、

上記複数のサブブロック毎に時間軸上で信号の実効値を求め、この実効値の標準偏差と平均値とに基づいてサブブロック毎の実効値の分布を求める工程と、

上記1ブロックの信号の周波数軸上のエネルギー分布を求める工程と、

上記1ブロックの信号のレベルを求める工程と、

上記複数のサブブロック毎の実効値の分布と上記1ブロックの信号の周波数軸上のエネルギー分布と上記1ブロックの信号のレベルとの少なくとも2つに応じて有声音か否かを判別する工程とを有することを特徴とする有声音判別方法。

【請求項7】 上記複数のサブブロック毎の実効値の分布と上記1ブロックの信号の周波数軸上のエネルギー分布と上記1ブロックの信号のレベルとの少なくとも1つの時間的な変化をトラッキングし、その結果に基づいて有声音か否かを判別することを特徴とする請求項6記載の有声音判別方法。

【請求項8】 上記1ブロックの信号について複数の周波数バンド毎に有声音／無声音の識別フラグを設定する際に、上記有声音判別工程において否と判別されたブロックは、全てのバンドを無声音フラグとすることを特徴とする請求項6記載の有声音判別方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、音声信号から有声音を雑音又は無声音と区別して判別する有声音判別方法に関する。

【0002】

【従来の技術】音声は音の性質として有声音と無声音に区別される。有声音は声帯振動を伴う音声で周期的な振動として観測される。無声音は声帯振動を伴わない音声で非周期的な音として観測される。通常の音声では大部分が有声音であり、無声音は無声子音と呼ばれる特殊な子音のみである。有声音の周期は声帯振動の周期で決まり、これをピッチ周期、その逆数をピッチ周波数という。これらピッチ周期及びピッチ周波数（以下、ピッチとした場合はピッチ周期を指す）は声の高低やイントネーションを決める重要な要因である。したがって、上記ピッチをどれだけ正確に捉えるかが音声の音質を左右する。しかし、上記ピッチを捉える場合には、上記音声の周囲にある雑音いわゆる背景雑音や量子化の際の量子化雑音を考慮しなければならない。これらの雑音又は無声

3

音と有声音を区別することが音声信号を符号化する場合に重要となる。

【0003】上記音声信号の符号化の具体的な例としては、MBE (Multiband Excitation: マルチバンド励起) 符号化、SBE (Singleband Excitation: シングルバンド励起) 符号化、ハーモニック (Harmonic) 符号化、SBC (Sub-band Coding: 帯域分割符号化)、LPC (Linear Predictive Coding: 線形予測符号化)、あるいはDCT (離散コサイン変換)、MDCT (モデファイドDCT)、FFT (高速フーリエ変換) 等がある。

【0004】例えば、上記MBE符号化においては、入力音声信号波形からピッチを抽出する場合、明確なピッチが表れない場合でもピッチの軌跡を捉えやすくしていた。そして、復号化側 (合成側) は、上記ピッチを基に余弦波 (cosin) 波合成により時間軸上の有声音波形を合成し、別途合成される時間軸上の無声音波形と加算合成し出力する。

【0005】

【発明が解決しようとする課題】ところで、ピッチを捉えやすくとすると上記背景雑音等の部分で本来のピッチでない間違ったピッチを捉えてしまう場合がある。もし、上記MBE符号化で間違ったピッチを捉えてしまうと、合成側では、その間違ったピッチの所で各cosin波のピークが重なるようにcosin波合成を行ってしまう。すなわち、誤って捉えたピッチ周期毎に有声音の合成で行っているような固定位相 (0位相又は $\pi/2$ 位相) の加算で各cosin波を合成し、ピッチが得られない筈の背景雑音等を周期性を持つインパルス波形として合成する。つまり、本来、時間軸上で散らばっているべき背景雑音等の振幅の強度があるフレームの1部分に周期性を持ちながら集中してしまい、非常に耳障りな異音を再生してしまうことになる。

【0006】本発明は、上記実情に鑑みてなされたものであり、有声音を雑音又は無声音と区別し確実に判別でき、合成側に対しては異音の発生を抑えさせることができる有声音判別方法の提供を目的とする。

【0007】

【課題を解決するための手段】本発明に係る有声音判別方法は、入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、1ブロックの信号を複数のサブブロックに分割する工程と、上記複数のサブブロック毎に信号の統計的な性質を求める工程と、上記統計的な性質の時間軸上での偏りに応じて有声音か否かを判別する工程とを有することを特徴として上記課題を解決することができる。

【0008】ここで、上記信号の統計的な性質には、各サブブロック毎の信号のピーク値、実効値又は標準偏差を用いることができる。

4

【0009】他の発明に係る有声音判別方法として、入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、1ブロックの信号の周波数軸上のエネルギー分布を求める工程と、上記1ブロックの信号のレベルを求める工程と、上記1ブロックの信号の周波数軸上のエネルギー分布と信号レベルとに応じて有声音か否かを判別する工程とを有することを特徴として上記課題を解決することができる。

10 【0010】ここで、上記各サブブロック毎の信号のピーク値、実効値又は標準偏差という統計的な性質と上記1ブロックの信号の周波数軸上のエネルギー分布とに応じて又は上記各サブブロック毎の信号のピーク値、実効値又は標準偏差という統計的な性質と上記1ブロックの信号のレベルとに応じて有声音か否かを判別してもよい。

【0011】さらに他の発明に係る有声音判別方法として、入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、1ブロックの信号を複数のサブブロックに分割する工程と、上記複数のサブブロック毎に時間軸上で信号のピーク値、実効値又は標準偏差を求める工程と、上記1ブロックの信号の周波数軸上のエネルギー分布を求める工程と、上記複数のサブブロック毎の信号のピーク値、実効値又は標準偏差と上記1ブロックの信号の周波数軸上のエネルギー分布と上記1ブロックの信号のレベルとに応じて有声音か否かを判別する工程とを有することを特徴として上記課題を解決することができる。

30 【0012】またさらに他の発明に係る有声音判別方法として、入力された音声信号をブロック単位で分割して各ブロック毎に有声音か否かの判別を行う有声音判別方法において、1ブロックの信号を複数のサブブロックに分割する工程と、上記複数のサブブロック毎に時間軸上で信号の実効値を求め、この実効値の標準偏差と平均値とに基づいてサブブロック毎の実効値の分布を求める工程と、上記1ブロックの信号の周波数軸上のエネルギー分布を求める工程と、上記複数のサブブロック毎の実効値の分布と上記1ブロックの信号の周波数軸上のエネルギー分布と上記1ブロックの信号のレベルとの少なくとも2つに応じて有声音か否かを判別する工程とを有することを特徴としている。

40 【0013】ここでいう有声音か否かの判別とは、有声音か雑音又は無声音かを判別することであり、有声音を確実に判別すると共に雑音又は無声音も確実に判別できる。つまり、入力音声信号から雑音 (背景雑音) 又は無声音を判別することもできる。このようなときには、例えば、強制的に入力音声信号の全帯域を無声音とする
50 と、合成側での異音の発生を抑えることができる。

【0014】

【作用】有声音と雑音又は無声音の統計的な性質の時間軸上で偏りが異なるため、入力音声信号が有声音か雑音又は無声音であるかを判別することができる。

【0015】

【実施例】以下、本発明に係る有声音判別方法の実施例について、図面を参照しながら説明する。図1は、本発明の第1の実施例となる有声音判別方法を説明するための有声音判別装置の概略構成を示している。この第1の実施例は、音声の1ブロックの信号をさらに分割したサブブロック毎の信号の統計的な性質の時間軸上での偏りに応じて有声音か否かを判別する。

【0016】図1において、入力端子11には、図示しないHPF（ハイパスフィルタ）等のフィルタによりいわゆるDC（直流）オフセット分の除去や帯域制限（例えば200～3400Hzに制限）のための少なくとも低域成分（200Hz以下）の除去が行われた音声の信号が供給される。この信号は、窓かけ処理部12に送られる。この窓かけ処理部12では1ブロックNサンプル

（例えばN=256）に対して方形窓をかけ、この1ブロックを1フレームLサンプル（例えばL=160）の間隔で時間軸方向に順次移動させており、各ブロック間のオーバーラップはN-Lサンプル（96サンプル）となっている。上記窓かけ処理部12からのNサンプルのブロックの信号は、サブブロック分割部13に供給される。このサブブロック分割部13は、上記窓かけ処理部12で分割された1ブロックの信号をさらに細分割する。そして、得られたサブブロック毎の信号は、統計的性質検出部14に供給される。この統計的性質検出部14は、本第1の実施例の場合、標準偏差又は実効値情報検出部15及びピーク値情報検出部16からなる。上記標準偏差又は実効値情報検出部15で得られた標準偏差又は実効値情報は、標準偏差又は実効値偏在検出部17に供給される。この標準偏差又は実効値偏在検出部17は、標準偏差又は実効値情報から時間軸上での偏りを検出する。そして、この時間軸上での標準偏差又は実効値の偏在情報は、判断部18に供給される。この判断部18は、時間軸上での標準偏差又は実効値の偏在情報を例えば所定の閾値と比較することによりサブブロック毎の信号が有声音であるか否かを判断し、その情報を出力端子20から導出する。一方、上記ピーク値情報検出部16で得られたピーク値情報は、ピーク値偏在検出部19に供給される。このピーク値偏在検出部19は、上記ピーク値情報から時間軸上での信号のピーク値の偏りを検出する。そして、この時間軸上での信号のピーク値の偏在情報は、判断部18に供給される。この判断部18は、上記時間軸上での信号のピーク値の偏在情報を例えば所定の閾値と比較することによりサブブロック毎の信号が有声音であるか否かを判断し、その判断情報を出力端子

20から導出する。

【0017】次に、本第1の実施例で統計的性質として用いられる各サブブロック毎の信号のピーク値情報、標準偏差又は実効値情報の検出とそれらの時間軸上での偏在の検出について説明する。

【0018】ここで、上記各サブブロック毎の信号のピーク値、標準偏差又は実効値を本第1の実施例で用いるのは、有声音と雑音又は無声音の信号のピーク値、標準偏差又は実効値が時間軸上で著しく異なるためである。例えば、図2のAに示すような音声の母音（有声音）と図2のCに示すような雑音又は子音（無声音）を比較する。母音の振幅のピークの並びは、図2のAのように時間軸上で偏りながらも規則的であるのに対し、雑音又は子音の振幅のピークの並びは時間軸上で一様（フラット）であるが不規則である。また、母音の標準偏差又は実効値も、図2のBに示すように時間軸上で偏っているのに対し、雑音又は子音の標準偏差又は実効値は、図2のDに示すように時間軸上でフラットである。

【0019】先ず、信号の上記各サブブロック毎の標準偏差又は実効値情報を検出する標準偏差又は実効値情報検出部15と該標準偏差又は実効値情報の時間軸上での偏在の検出について説明する。この標準偏差又は実効値情報検出部15は、図3に示すように入力端子21からのサブブロック毎の信号から標準偏差又は実効値を算出する標準偏差又は実効値算出部22と、該標準偏差又は実効値から相加平均を算出する相加平均算出部23と、上記標準偏差又は実効値から相乗平均値を算出する相乗平均算出部24とからなる。そして、上記相加平均値と相乗平均値より時間軸上での偏在情報を標準偏差又は実効値偏在検出部17が検出し、判断部18が該偏在情報からサブブロック毎の音声信号が有声音か否かを判断し、その判断情報が出力端子20から導出される。

【0020】上記エネルギーの分散から有声音か否かを判断する原理を図1と図3を用いて説明する。上記窓かけ処理部12で方形窓をかけることにより切り出される1ブロックのサンプル数Nを256サンプルとし、入力サンプル列を $x(n)$ とする。この1ブロック（256サンプル）を上記サブブロック分割部13により8サンプル毎に分割する。するとサブブロック長 $B_1=8$ のサブブロックが N/B_1 （ $256/8=32$ ）個上記1ブロックの中に存在することになる。この32個のサブブロック毎の時間軸上データは、上記標準偏差又は実効値情報検出部15の例えば標準偏差又は実効値算出部22に供給される。

【0021】この標準偏差又は実効値算出部22は、上記32個のサブブロック毎に時間軸上データの例えば標準偏差 $\sigma_a(i)$ として、

【0022】

【数1】

$$\sigma_a(i) = \sqrt{\frac{1}{B_i} \sum_{n=k}^{k+B_i-1} (x(n) - \underline{x})^2} \quad 0 \leq i < N/B_i$$

$$k = i \times B_i \quad 0 \leq i < N/B_i$$

... (1)

【0023】で示される(1)式により算出した値を出力する。ここで*i*はサブブロックのインデックスであり、*k*はサンプル数である。また、 \underline{x} は1ブロック当たりの入力サンプルの平均値である。この平均値 \underline{x} は、1

【0024】また、上記サブブロック毎の実効値は、上記(1)式中の $(x(n) - \underline{x})^2$ の代わりに、各サンプル*x*について上記1ブロック内のサンプルの平均値 \underline{x} との差をとらない $(x(n))^2$ を用いたものであり、

$$a_{v:add} = \frac{1}{N/B_i} \sum_{i=0}^{N/B_i-1} \sigma_a(i) \quad \dots (2)$$

$$a_{v:mpy} = \left\{ \prod_{i=0}^{N/B_i-1} \sigma_a(i) \right\}^{1/N/B_i} \quad \dots (3)$$

【0027】で示される(2)及び(3)式により算出する。これらの(1)式～(3)式では標準偏差についてのみ例示しているが、実効値の場合も同様であることは勿論である。

【0028】上記(2)及び(3)式により算出された

$$p_f = a_{v:add} / a_{v:mpy}$$

で求める。この比率 p_f は、時間軸上の標準偏差の偏在を表す偏在情報である。この偏在情報(比率) p_f は、判断部18に供給され、該判断部18では、例えば、上記偏在情報 p_f を閾値 p_{thf} と比較し有声音か否かの判断を行う。例えば、上記閾値 p_{thf} を1.1に設定しておき、上記偏在情報 p_f が該閾値 p_{thf} より大きいと標準偏差又は実効値の偏りが大きいと判断し有声音とする。一方、上記分散情報 p_f が該閾値 p_{thf} より小さいと標準偏差又は実効値の偏りが小さい(フラットである)と判断し有声音でない(雑音又は無声音である)とする。

【0029】次に、ピーク値情報を検出するピーク値情報検出部16と該ピーク値の時間軸上での偏在の検出について説明する。このピーク値情報検出部16は、図4に示すように入力端子21からのサブブロック毎の信号からピーク値を検出するピーク値検出部26と、このピーク値検出部26からのピーク値の平均値を算出する平均ピーク値算出部27と、入力端子25を介して供給されるブロック毎の信号から標準偏差値を算出する標準偏

いわゆるrms (root meansquare、自乗平均の平方根)とも称されるものである。

【0025】上記標準偏差 $\sigma_a(i)$ は、時間軸上での分散を調べるために上記相加平均算出部23及び相乗平均算出部24に供給される。上記相加平均算出部23及び相乗平均算出部24は、相加平均値 $a_{v:add}$ 及び相乗平均値 $a_{v:mpy}$ を、

【0026】

【数2】

相加平均値 $a_{v:add}$ 及び相乗平均値 $a_{v:mpy}$ は、上記標準偏差又は実効値偏在検出部17に供給される。この標準偏差又は実効値偏在検出部17は、上記相加平均値 $a_{v:add}$ と相乗平均値 $a_{v:mpy}$ とから比率 p_f を、

... (4)

差算出部28とからなる。そして、上記ピーク値偏在検出部19が上記平均ピーク値算出部27からの平均ピーク値を上記標準偏差算出部28からのブロック毎の標準偏差値で除算し、時間軸上での平均ピーク値の偏在を検出する。この平均ピーク値偏在情報は、判断部18に供給される。この判断部18が該平均ピーク値偏在情報を基にサブブロック毎の音声信号が有声音か否かを判断し、該判断情報が出力端子20から導出される。

【0030】上記ピーク値情報から有声音か否かを判断する原理を図1と図4を用いて説明する。上記ピーク値検出部26には、上記窓かけ処理部12、サブブロック分割部13及び入力端子21を介してサブブロック長 B_1 (例えば8)のサブブロック分の信号が N/B_1 (256/8=32)個供給される。このピーク値検出部26は、例えば32個分のサブブロック毎のピーク値 $P(i)$ を、

【0031】

【数3】

9

$$P(i) = \text{MAX}_{k \leq n \leq k+B_i-1} (|x(n)|)$$

$$k = i \times B_i$$

10

$$0 \leq i < N/B_i$$

... (5)

【0032】で示される(5)式の条件で検出する。ここで*i*はサブブロックのインデックスであり、*k*はサンプル数である。また、MAXは最大値を求める関数である。

$$\bar{P} = \frac{1}{N/B_i} \sum_{i=0}^{N/B_i-1} P(i)$$

【0035】で示される(6)式により算出する。

【0036】また、上記標準偏差算出部28は、ブロック毎の標準偏差値 $\sigma_b(i)$ を、

$$\sigma_b(i) = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} (x(n) - \bar{x})^2} \quad 0 \leq i < N$$

... (7)

【0038】で求める。そして、上記ピーク値偏在検出部19は、ピーク値偏在情報 P_n を上記平均ピーク値 \bar{P}

$$P_n = \bar{P} / \sigma_b(i)$$

のように算出する。なお、上記標準偏差算出部28の代わりに、実効値(rms値)を算出する実効値算出部を用いてもよい。

【0039】上記(8)式により算出されたピーク値偏在情報 P_n は、時間軸上でのピーク値の偏在の度合いを示すもので、上記判断部18に供給される。そして、上記判断部18は、例えば、上記ピーク値偏在情報 P_n を閾値 P_{thn} と比較し有声音か否かの判断を行う。例えば、上記、ピーク値偏在情報 P_n が該閾値 P_{thn} より大きいとピーク値の時間軸上での偏りが大きいと判断し有声音とする。一方、上記ピーク値偏在情報 P_n が閾値 P_{thn} より小さいとピーク値の偏りが小さいと判断し有声音でない(雑音又は無声音である)とする。

【0040】以上により、本発明に係る有声音判別方法の第1の実施例は、各サブブロック毎の信号のピーク値、実効値又は標準偏差のような統計的性質の時間軸上での偏りに応じて有声音か否かを判別することができる。

【0041】次に図5は、本発明の第2の実施例としての有声音判別方法を説明するための有声音判別装置の概略構成を示す図である。この第2の実施例は、音声の1ブロックの信号の周波数軸上のエネルギーの分布とレベルとから有声音か否かを判別する。

【0042】この第2の実施例は、有声音のエネルギー分布が周波数軸上の低域側に集中し、雑音又は無声音のエネルギー分布が周波数軸上の高域側に集中する傾向を用いている。

【0043】この図5において、入力端子31には、図

【0033】そして、上記平均ピーク値算出部27が上記ピーク値 $P(i)$ から平均ピーク値 \bar{P} を、

【0034】

【数4】

... (6)

【0037】

【数5】

20 と上記標準偏差値 $\sigma_b(i)$ とから、

... (8)

示しないHPF(ハイパスフィルタ)等のフィルタによりいわゆるDC(直流)オフセット分の除去や帯域制限(例えば200~3400Hzに制限)のための少なくとも低域成分(200Hz以下)の除去が行われた音声の信号が供給される。この信号は、窓かけ処理部32に送られる。この窓かけ処理部32では1ブロックNサンプル(例えばN=256)に対して例えばハミング窓をかけ、この1ブロックを1フレームLサンプル(例えばL=160)の間隔で時間軸方向に順次移動させており、各ブロック間のオーバーラップはN-L(96サンプル)となっている。この窓かけ処理部32でNサンプルのブロックとされた信号は、直交変換部33に供給される。この直交変換部33は、例えば1ブロック256サンプルのサンプル列に対して1792サンプル分の0データを付加して(いわゆる0詰めして)2048サンプルとし、この2048サンプルの時間軸データ列に対して、FFT(高速フーリエ変換)等の直交変換処理を施し、周波数軸データ列に変換する。この直交変換部33からの周波数軸上のデータは、エネルギー検出部34に供給される。このエネルギー検出部34は、供給された周波数軸上データを低域側と高域側に分け、それぞれ低域側エネルギー検出部34aと高域側エネルギー検出部34bによりエネルギーを検出する。この低域側エネルギー検出部34a及び高域側エネルギー検出部34bにより検出された低域側エネルギー検出値及び高域側エネルギー検出値は、エネルギー分布算出部35に供給され、比率(エネルギー分布情報)が求められる。このエネルギー分布算出部35により求められたエネルギー分

40

50

布情報は、判断部37に供給される。また、上記低域側エネルギー検出値と高域側エネルギー検出値は、信号レベル算出部36に供給され、1サンプル当たりの信号のレベルが計算される。この信号レベル算出部36によって算出された信号レベル情報は、上記判断部37に供給される。上記判断部37は、上記エネルギー分布情報及び信号レベル情報を基に入力音声信号が有声音であるかを判断し、判断情報を出力端子38から導出する。

【0044】以下に、この第2の実施例の動作を説明する。上記窓かけ処理部32でハミング窓をかけることに

$$a_m(j) = \sqrt{R_e^2(j) + I_m^2(j)} \quad 10$$

【0046】により求められる。この(9)式で $R_e(j)$ は実数部を表し、 $I_m(j)$ は虚数部を表す。また、 j はサンプル数で0以上 $N/2$ ($=128$ サンプル) 未満の範囲にある。

【0047】上記エネルギー検出部34の低域側エネルギー

$$S_L = \sum_{j=0}^{N/4-1} a_m^2(j)$$

$$S_H = \sum_{j=N/4}^{N/2-1} a_m^2(j)$$

【0049】で示される(10)式及び(11)式により求められる。ここでいう低域側は0~2 KHz、高域側は2~3.4 KHzの周波数帯である。上記(10)、(11)式により算出された低域側エネルギー S_L 及び高域側エネルギー S_H は

$$f_b = S_L / S_H$$

となる。

【0050】この周波数軸上でのエネルギー分布情報 f_b は、判断部37に供給される。この判断部37は、上記エネルギー分布情報 f_b を例えば閾値 f_{thb} と比較し有声音か否かの判断を行う。例えば上記閾値 f_{thb} を1.5に設定しておき上記エネルギー分布情報 f_b が該閾値 f_{thb} より小さいときは高域側にエネルギーが集中して有声音でない(雑音又は無声音である)確率が高い

$$l_o = \sqrt{\frac{S_L + S_H}{N/2}}$$

【0053】で示される(13)式から求める。この平均レベル情報 l_a も判断部37に供給される。この判断部37は、上記平均レベル情報 l_a を例えば閾値 l_{tha} と比較し有声音か否かの判断を行う。例えば上記閾値 l_{tha} を550に設定しておき上記平均レベル情報 l_a が該閾値 l_{tha} より小さいときは有声音でない(雑音又は無声音である)確率が高いと判断することになる。

【0054】上記判断部37は、上記エネルギー分布情報

より切り出される1ブロックのサンプル数 N を256サンプルとし、入力サンプル列を $x(n)$ とする。この1ブロック(256サンプル)の時間軸上のデータは、上記直交変換部33により1ブロックの周波数軸上のデータに変換される。この1ブロックの周波数軸上のデータは、上記エネルギー検出部34に供給され、振幅 $a_m(j)$ が、

【0045】

【数6】

$$\dots (9)$$

ギー検出部34a及び高域側エネルギー検出部34bでは、上記(9)式に示された振幅 $a_m(j)$ から、低域側エネルギー S_L 及び高域側エネルギー S_H 及びを、

【0048】

【数7】

$$\dots (10)$$

$$\dots (11)$$

上記分布算出部35に供給され、その比率 S_L / S_H により周波数軸上でのエネルギーの分布のバランス情報(エネルギー分布情報) f_b が求められる。すなわち、

$$\dots (12)$$

と判断することになる。

【0051】また、上記低域側エネルギー S_L 及び高域側エネルギー S_H は、上記信号レベル算出部36に供給される。この信号レベル算出部36は、上記低域側エネルギー S_L 及び高域側エネルギー S_H とを用いて、信号の平均レベル l_a 情報を、

【0052】

【数8】

$$\dots (13)$$

報 f_b と平均レベル情報 l_a の内のどちらか一つの情報からでも上述したように有声音か否かの判断が可能であるが、両方の情報を用いれば判断の信頼度は高くなる。すなわち、

$$f_b < f_{thb} \quad \text{かつ} \quad l_a < l_{tha}$$

のとき有声音でないという信頼度の高い判断ができる。

そして、出力端子38から該判断情報を導出する。

【0055】ここで、この第2の実施例での上記エネル

ギー分布情報 f_b と平均レベル情報 l_a を別々に、上述した第1の実施例での時間軸上の標準偏差又は実効値の偏在情報ある比率（偏在情報） p_f と組み合わせて有声音か否かの判断を行うこともできる。すなわち、 $p_f < p_{thf}$ かつ $f_b < f_{thb}$ 又は $p_f < p_{thf}$ かつ $l_a < l_{tha}$ のとき有声音でないという信頼度の高い判断を行うことができる。

【0056】以上により、この第2の実施例は、有声音のエネルギー分布が周波数軸上の低域側に集中し、雑音又は無声音のエネルギー分布が周波数軸上の高域側に集中する傾向を用いて有声音か否かを判別することができる。

【0057】次に図6は、本発明の第3の実施例としての有声音判別方法を説明するための有声音判別装置の概略構成を示す図である。

【0058】この図6において、入力端子41には、少なくとも低域成分（200Hz以下）が除去され、方形窓により1ブロックNサンプル（例えばN=256）で窓かけ処理されて時間軸方向に移動され、さらに1ブロックが細分割されたサブブロック毎の信号が供給される。このサブブロック毎の信号から上記統計的性質検出部14が統計的性質を検出する。そして上記第1の実施例で説明したような偏在検出部17又は19が上記統計的性質から統計的性質の時間軸上での偏りを検出する。この偏在検出部17又は19からの偏在情報は、判断部39に供給される。また、入力端子42には、少なくとも低域成分（200Hz以下）が除去され、ハミング窓により1ブロックNサンプル（例えばN=256）で窓かけ処理されて時間軸方向に移動され、さらに直交変換により周波数軸上に変換されたデータが供給される。この周波数軸上に変換されたデータは、上記エネルギー検出部34に供給される。このエネルギー検出部34により検出された高域側エネルギー検出値と低域側エネルギー検出値は、エネルギー分布算出部35に供給される。このエネルギー分布計算部35により求められたエネルギー分布情報は、判断部39に供給される。さらに、上記高域側エネルギー検出値と低域側エネルギー検出値は、信号レベル算出部36に供給され、1サンプル当たりの信号のレベルが計算される。この信号レベル計算部36によって計算された信号レベル情報は、上記判断部39に供給される。上記判断部39には、上記偏在情報、エネルギー分布情報及び信号レベル情報が供給される。これらの情報により判断部39は、入力音声信号が有声音であるか否かを判断する。そして、出力端子43から該判断情報を導出する。

【0059】以下に、この第3の実施例の動作を説明する。この第3の実施例は、上記偏在検出部17、19からの各サブフレーム毎の信号の偏向情報 p_f 、上記分布算出部35からのエネルギー分布情報 f_b 及び上記信号

レベル算出部36からの平均レベル情報 l_a を用いて上記判断部39で有声音か否かの判断を行うものである。例えば、

$$p_f < p_{thf} \quad \text{かつ} \quad f_b < f_{thb} \quad \text{かつ} \quad l_a < l_{tha}$$

のとき有声音でないという信頼度の高い判断を行う。

【0060】以上により、この第3の実施例は、統計的性質の時間軸上での偏在情報、エネルギー分布情報及び平均レベル情報とに応じて有声音か否かを判断する。

10 【0061】なお、本発明の上記実施例に係る有声音判別方法は、上記具体例にのみ限定されるものでないことはいふまでもない。例えば、各サブフレーム毎の信号の偏在情報 p_f を用いて有声音を判別する場合には、その時間変化を追い例えば5フレーム連続して

$$p_f < p_{thf} \quad (p_{thf} = 1.1)$$

のときに限りフラットとみなしフラグ P_{fs} を1とする。

一方、5フレームの内1フレームでも、

$$p_f \geq p_{thf}$$

となったら、上記フラグ P_{fs} を0とする。そして、

$$20 \quad f_b < f_{bt} \quad \text{かつ} \quad P_{fs} = 1 \quad \text{かつ} \quad l_a < l_{tha}$$

のときに有声音でないという信頼度の非常に高い判断を行うことができる。

【0062】そして、本発明に係る有声音判別方法により、有声音でない、すなわち、背景雑音又は子音と判断されたときには、入力音声信号の1ブロックを全て強制的に無声音とすることにより、MBE等のボコーダの合成側での異音の発生を防ぐことができる。

30 【0063】次に、本発明に係る有声音判別方法の第4の実施例について、図7及び図8を参照しながら説明する。上述した第1の実施例においては、信号の上記サブブロック毎の標準偏差や実効値（rms値）のデータの分布を調べるために、標準偏差や実効値の各データの相加平均と相乗平均との比率を求めているが、上記相乗平均をとるためには、上記1フレーム内のサブブロックの個数（例えば32個）のデータの乗算と32乗根の演算とが必要とされる。この場合、先に32個のデータを乗算するとオーバーフロー（桁あふれ）が生ずるため、先に各データのそれぞれ32乗根をとった後に乗算を行うような工夫が必要とされる。このとき、32個の各データ毎に32回の32乗根演算が必要となり、多くの演算量が要求されることになる。

【0064】そこで、この第4の実施例においては、上記32個の各サブブロック毎の実効値（rms値）のフレーム内での標準偏差 σ_{rms} と平均値 \bar{rms} とを求め、これらの値に応じて（例えばこれらの値の比率に応じて）実効値 rms の分布を検出している。すなわち、上記各サブブロック毎の実効値 rms 、この rms のフレーム内の標準偏差 σ_{rms} 及び平均値 \bar{rms} は、

【0065】

【数9】

15

$$rms(i) = \sqrt{\frac{1}{B_L} \sum_{j=0}^{B_L-1} x^2(i \cdot B_L + j)} \quad 0 \leq i < B_N (=32) \quad \dots (14)$$

$$\underline{rms} = \frac{1}{B_N} \sum_{i=0}^{B_N-1} rms(i) \quad (B_N = 32) \quad \dots (15)$$

$$\sigma_{rms} = \sqrt{\frac{1}{B_N} \sum_{i=0}^{B_N-1} (rms(i) - \underline{rms})^2} \quad \dots (16)$$

【0066】と表せる。これらの式で、 i は上記サブブロックのインデックス（例えば $i=0 \sim 31$ ）、 B_L はサブブロック内のサンプル数（サブブロック長、例えば $B_L=8$ ）、 B_N は1フレーム内のサブブロックの個数（例えば $B_N=32$ ）をそれぞれ示し、1フレーム内のサンプル数 N を例えば256としている。

$$\sigma_m = \sigma_{rms} / \underline{rms}$$

となる。この σ_m は、有声部では大きな値となり、無声部又は背景雑音部分では小さな値となる。この σ_m が閾値 σ_{th} より大きいときは有声とみなし、閾値 σ_{th} より小さいときは無声又は背景雑音の可能性ありとして、他の条件（信号レベルやスペクトルの傾き）のチェックを行う。なお、上記閾値 σ_{th} の具体的な値としては、 $\sigma_{th}=0.4$ が挙げられる。

【0068】以上のような時間軸上のエネルギー分布の分析処理は、図8のAに示すような音声の母音部と図8のBに示すようなノイズ又は音声の子音部とで、上記サブフレーム毎の短時間実効値（ rms 値）の分布に違いが見られることに着目したものである。すなわち、図8のAの母音部での上記短時間 rms 値の分布（曲線b参照）には大きな偏りがあるのに対して、図8のBのノイズ又は子音部での短時間 rms 値の分布（曲線b）はほぼフラットである。なお、図8のA、Bの各曲線aは信号波形（サンプル値）を示している。このような短時間 rms 値の分布を調べるために、本実施例では、短時間 rms 値のフレーム内の標準偏差 σ_{rms} と平均値 \underline{rms} との比率、すなわち上記正規化（ノーマライズ）された標準偏差を σ_m を用いているわけである。

【0069】この時間軸上のエネルギー分布の分析処理のための構成については、図7の入力端子51からの入力データを、実効値算出部61に送って上記サブブロック毎の実効値 $rms(i)$ を求め、平均値及び標準偏差算出部62に送って上記平均値 \underline{rms} 及び標準偏差 σ_{rms} を求めた後、正規化標準偏差算出部63に送って上記正規化した標準偏差 σ_m を求めている。この正規化標準偏差 σ_m は、ノイズ又は無声区間判別部64に送ってい

16

【0067】上記(16)式の標準偏差 σ_{rms} は、信号レベルが大きくなるとそれだけで大きくなってしまいうので、上記(15)式の平均値 \underline{rms} で割り込んで正規化（ノーマライズ）する。この正規化（ノーマライズ）した標準偏差を σ_m とすると、

$$\dots (17)$$

る。

【0070】次に、スペクトルの傾きのチェックについて説明する。通常、有声音部分では、周波数軸上で低域にエネルギーが集中する。これに対して無声部又は背景雑音部では高域側にエネルギーが集中しやすい。そこで、高域側と低域側のエネルギーの比をとって、その値を雑音部か否かの評価尺度の1つとして使用する。すなわち、図7の入力端子51からの1ブロック（1フレーム）内の $x(n)$ （ $0 \leq n < N$ 、 $N=256$ ）に対して、窓かけ処理部52にて適当な窓（例えばハミング窓）をかけ、FFT（高速フーリエ変換）部53でFFT処理を行って得た結果を、

$$Re(j) \quad (0 \leq j < N/2)$$

$$Im(j) \quad (0 \leq j < N/2)$$

とする。ただし、 $Re(j)$ はFFT係数の実部、 $Im(j)$ は同虚部である。また、 $N/2$ は規格化周波数の π に相当し、実周波数の4kHz（ $x(n)$ は8kHzサンプリングのデータなので）に当たる。

【0071】上記FFT処理結果は、振幅算出部54に送って振幅 $a_m(j)$ を求めている。この振幅算出部54は、上記第2の実施例のエネルギー検出部34と同様な処理を行う部分であり、上記(9)式の演算が行われる。次に、この演算結果である振幅 $a_m(j)$ が S_L 、 S_H 、 f_b 算出部55に送られ、この算出部55において、上記エネルギー検出部34内の低域側、高域側の各エネルギー検出部34a、34bでの演算、すなわち上記(10)式による低域側エネルギー S_L の演算、及び上記(11)式による高域側エネルギー S_H の演算が行われ、さらにこれらの比率であるエネルギーバランスを示

20

30

40

50

17

すパラメータ f_b ($=S_L/S_H$ 、上記(12)式参照)を求めている。この値が小さいときは高域側にエネルギーが片寄っていてノイズ又は子音である可能性が高い。このパラメータ f_b を上記ノイズ又は無声区間判別部64に送っている。

【0072】次に、上記第2の実施例の信号レベル算出部36に相当する信号パワー算出部56において、上記(13)式に示す信号の平均レベルあるいはパワー l_a を算出している。この信号レベルあるいは信号パワー l_a も上記ノイズ又は無声区間判別部64に送っている。

【0073】ノイズ又は無声区間判別部64においては、上記各算出された値 σ_m 、 f_b 、 l_a に基づいてノイズ又は無声区間を判別する。この判別ための処理を $F(\cdot)$ と定義するとき、 $F(\sigma_m, f_b, l_a)$ の関数の具体例として次のようなものが挙げられる。

【0074】先ず、第1の具体例として、

$$f_b < f_{bth} \text{ かつ } \sigma_m < \sigma_{mth} \text{ かつ } l_a < l_{ath}$$

ただし、 f_{bth} 、 σ_{mth} 、 l_{ath} はいずれも閾値の条件とすることが考えられ、この条件が満足されるとき、ノイズと判断し、全バンドUV(無声音)とする。ここで、各閾値の具体的な値としては、 $f_{bth} = 1.5$ 、 $\sigma_{mth} = 0.4$ 、 $l_{ath} = 550$ が挙げられる。

【0075】次に、第2の例として、上記正規化標準偏差 σ_m の信頼度を向上するために、もう少し長時間の σ_m を観測することも考えられる。具体的には、Mフレーム連続して $\sigma_m < \sigma_{mth}$ のときに限り、時間軸上のエネルギー分布がフラットであるとし、 σ_m 状態フラグ σ_{state} をセット ($\sigma_{state} = 1$) する。1フレームでも $\sigma_m \leq \sigma_{mth}$ が出現したときには、上記 σ_m 状態フラグ σ_{state} をリセット ($\sigma_{state} = 0$) する。そして、

$$f_b < f_{bth} \text{ かつ } \sigma_{state} = 1 \text{ かつ } l_a < l_{ath}$$

のときにノイズあるいは無声と判断し、V/UVフラグをオールUVとする。

【0076】上記第2の例のように正規化標準偏差 σ_m の信頼度を高めた状態においては、信号レベル(信号パワー) l_a のチェックを不要としてもよい。この場合の関数 $F(\cdot)$ としては、

$$f_b < f_{bth} \text{ かつ } \sigma_{state} = 1$$

のときに、無声又はノイズと判断すればよい。

【0077】以上説明したような第4の実施例によれば、DSPへのインプリメントが可能な程度の少ない演算量で、正確にノイズ(背景雑音)区間や無声区間を検

$$x_w(k, q) = x(q) w(kL - q)$$

となる。この(18)式において、 k はブロック番号を、 q はデータの時間インデックス(サンプル番号)を表し、処理前の入力信号の q 番目のデータ $x(q)$ に対して第 k ブロックの窓(ウィンドウ)関数 $w(kL - q)$ により窓

$$w_r(r) = 1 \quad 0 \leq r < N$$

$$= 0 \quad r < 0, N \leq r$$

18

出することが可能となり、背景雑音と判定された部分(フレーム)は強制的に全バンドをUVとすることで、背景雑音をエンコード/デコードすることによるうなり音のような異音の発生を抑えることが可能になる。

【0078】以下、本発明に係る有声音判別方法が適用可能な音声信号の合成分析符号化装置(いわゆるボコーダ)の一種のMBE(Multiband Excitation: マルチバンド励起)ボコーダの具体例について、図面を参照しながら説明する。このMBEボコーダは、D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, No. 8, pp. 1223-1235, Aug. 1988に開示されているものであり、従来のPARCOR(PARTIAL-CORrelation: 偏自己相関)ボコーダ等では、音声のモデル化の際に有声音区間と無声音区間とをブロックあるいはフレーム毎に切り換えていたのに対し、MBEボコーダでは、同時刻(同じブロックあるいはフレーム内)の周波数軸領域に有声音(Voiced)区間と無声音(Unvoiced)区間とが存在するという仮定でモデル化している。

【0079】図9は、上記MBEボコーダの実施例の全体の概略構成を示すブロック図である。この図9において、入力端子101には音声信号が供給されるようになっており、この入力音声信号は、HPF(ハイパスフィルタ)等のフィルタ102に送られて、いわゆるDC(直流)オフセット分の除去や帯域制限(例えば200~3400Hzに制限)のための少なくとも低域成分(200Hz以下)の除去が行われる。このフィルタ102を介して得られた信号は、ピッチ抽出部103及び窓かけ処理部104にそれぞれ送られる。ピッチ抽出部103では、入力音声信号データが所定サンプル数 N (例えば $N = 256$) 単位でブロック分割され(あるいは方形窓による切り出しが行われ)、このブロック内の音声信号についてのピッチ抽出が行われる。このような切り出しブロック(256サンプル)を、例えば図10のAに示すように L サンプル(例えば $L = 160$) のフレーム間隔で時間軸方向に移動させており、各ブロック間のオーバーラップは $N - L$ サンプル(例えば96サンプル)となっている。また、窓かけ処理部104では、1ブロック N サンプルに対して所定の窓関数、例えばハミング窓をかけ、この窓かけブロックを1フレーム L サンプルの間隔で時間軸方向に順次移動させている。

【0080】このような窓かけ処理を数式で表すと、

$$\dots (18)$$

かけ処理されることによりデータ $x_w(k, q)$ が得られることを示している。ピッチ抽出部103内での図10のAに示すような方形窓の場合の窓関数 $w_r(r)$ は、

$$\dots (19)$$

また、窓かけ処理部104での図10のBに示すような

$$w_h(r) = 0.54 - 0.46 \cos(2\pi r/(N-1))$$

$$= 0$$

である。このような窓関数 $w_r(r)$ あるいは $w_h(r)$ を用いるときの上記(18)式の窓関数 $w(r)$ ($=w(kL-q)$)の非零区間は、

$$0 \leq kL - q < N$$

これを变形して、

$$kL - N < q \leq kL$$

従って、例えば上記方形窓の場合に窓関数 $w_r(kL-q) = 1$ となるのは、図11に示すように、 $kL - N < q \leq kL$ のときとなる。また、上記(18)～(20)式は、長さ $N (=256)$ サンプルの窓が、 $L (=160)$ サンプルずつ前進してゆくことを示している。以下、上記(19)式、(20)式の各窓関数で切り出された各 N 点 ($0 \leq r < N$)の非零サンプル列を、それぞれ $x_{wr}(k, r)$ 、 $x_{wh}(k, r)$ と表すことにする。

【0081】窓かけ処理部104では、図12に示すように、上記(20)式のハミング窓がかけられた1ブロック256サンプルのサンプル列 $x_{wh}(k, r)$ に対して1792サンプル分の0データが付加されて(いわゆる0詰めされて)2048サンプルとされ、この2048サンプルの時間軸データ列に対して、直交変換部105により例えばFFT(高速フーリエ変換)等の直交変換処理が施される。

【0082】ピッチ抽出部103では、上記 $x_{wr}(k, r)$ のサンプル列(1ブロック N サンプル)に基づいてピッチ抽出が行われる。このピッチ抽出法には、時間波形の周期性や、スペクトルの周期的周波数構造や、自己相関関数を用いるもの等が知られているが、本実施例では、センタクリップ波形の自己相関法を採用している。このときのブロック内でのセンタクリップレベルについては、1ブロックにつき1つのクリップレベルを設定してもよいが、ブロックを細分割した各部(各サブブロック)の信号のピークレベル等を検出し、これらの各サブブロックのピークレベル等の差が大きいときに、ブロック内でクリップレベルを段階的にあるいは連続的に変化

$$S(j) = H(j) | E(j) |$$

と表現するようなモデルを想定している。ここで、 J は $\pi \omega_s = f_s / 2$ に対応し、サンプリング周波数 $f_s = 2\pi \omega_s$ が例えば8kHzのときには4kHzに対応する。上記(21)式中において、周波数軸上のスペクトルデータ $S(j)$ が図13のAに示すような波形のとき、 $H(j)$ は、図13のBに示すような元のスペクトルデータ $S(j)$ のスペクトル包絡線(エンベロープ)を示し、 $E(j)$ は、図13のCに示すような等レベルで周期的な励起信号(エキサイテーション)のスペクトルを示している。すなわち、FFTスペクトル $S(j)$ は、スペクトルエンベロープ $H(j)$ と励起信号のパワースペクトル $|E(j)|$

ハミング窓の場合の窓関数 $w_h(r)$ は、

$$0 \leq r < N \quad \dots (20)$$

$$r < 0, N \leq r$$

させるようにしている。このセンタクリップ波形の自己相関データのピーク位置に基づいてピーク周期を決めている。このとき、現在フレームに属する自己相関データ(自己相関は1ブロック N サンプルのデータを対象として求められる)から複数のピークを求めておき、これらの複数のピークの内の最大ピークが所定の閾値以上のときには該最大ピーク位置をピッチ周期とし、それ以外のときには、現在フレーム以外のフレーム、例えば前後のフレームで求められたピッチに対して所定の関係を満たすピッチ範囲内、例えば前フレームのピッチを中心として $\pm 20\%$ の範囲内にあるピークを求め、このピーク位置に基づいて現在フレームのピッチを決定するようにしている。このピッチ抽出部103ではオープンループによる比較的ラフなピッチのサーチが行われ、抽出されたピッチデータは高精度(ファイン)ピッチサーチ部106に送られて、クローズドループによる高精度のピッチサーチ(ピッチのファインサーチ)が行われる。

【0083】高精度(ファイン)ピッチサーチ部106には、ピッチ抽出部103で抽出された整数(インテジャ)値の粗(ラフ)ピッチデータと、直交変換部105により例えばFFTされた周波数軸上のデータとが供給されている。この高精度ピッチサーチ部106では、上記粗ピッチデータ値を中心に、0.2～0.5きざみで±数サンプルずつ振って、最適な小数点付き(フローティング)のファインピッチデータの値へ追い込む。このときのファインサーチの手法として、いわゆる合成による分析(Analysis by Synthesis)法を用い、合成されたパワースペクトルが原音のパワースペクトルに最も近くなるようにピッチを選んでいる。

【0084】このピッチのファインサーチについて説明する。先ず、上記MBEボコーダにおいては、上記FFT等により直交変換された周波数軸上のスペクトルデータとしての $S(j)$ を

$$0 < j < J \quad \dots (21)$$

(j) | との積としてモデル化される。

【0085】上記励起信号のパワースペクトル $|E(j)|$ は、上記ピッチに応じて決定される周波数軸上の波形の周期性(ピッチ構造)を考慮して、1つの帯域(バンド)の波形に相当するスペクトル波形を周波数軸上の各バンド毎に繰り返すように配列することにより形成される。この1バンド分の波形は、例えば上記図12に示すような256サンプルのハミング窓関数に1792サンプル分の0データを付加(0詰め)した波形を時間軸信号と見なしてFFTし、得られた周波数軸上のある帯域幅を持つインパルス波形を上記ピッチに応じて切り出す

ことにより形成することができる。

【0086】次に、上記ピッチに応じて分割された各バンド毎に、上記 $H(j)$ を代表させるような(各バンド毎のエラーを最小化するような)値(一種の振幅) $|A_m|$ を求める。ここで、例えば第 m バンド(第 m 高調波の

$$\varepsilon_m = \sum_{j=a_m}^{b_m} \{ |S(j)| - |A_m| |E(j)| \} \quad \dots (22)$$

【0088】で表せる。このエラー ε_m を最小化するよ
うな $|A_m|$ は、

$$\frac{\partial \varepsilon_m}{\partial |A_m|} = 2 \sum_{j=a_m}^{b_m} \{ |S(j)| - |A_m| |E(j)| \} |E(j)|$$

$$\therefore |A_m| = \frac{\sum_{j=a_m}^{b_m} |S(j)| |E(j)|}{\sum_{j=a_m}^{b_m} |E(j)|^2} \quad \dots (23)$$

【0090】となり、この(23)式の $|A_m|$ のとき、エラー ε_m を最小化する。このような振幅 $|A_m|$ を各バンド毎に求め、得られた各振幅 $|A_m|$ を用いて上記(22)式で定義された各バンド毎のエラー ε_m を求める。次に、このような各バンド毎のエラー ε_m の全バンドの総和値 $\sum \varepsilon_m$ を求める。さらに、このような全バンドのエラー総和値 $\sum \varepsilon_m$ を、いくつかの微小に異なるピッチについて求め、エラー総和値 $\sum \varepsilon_m$ が最小となるようなピッチを求める。

【0091】すなわち、上記ピッチ抽出部103で求められたラフピッチを中心として、例えば0.25きざみで上下に数種類ずつ用意する。これらの複数種類の微小に異なるピッチの各ピッチに対してそれぞれ上記エラー総和値 $\sum \varepsilon_m$ を求める。この場合、ピッチが定まるとバンド幅が決まり、上記(23)式より、周波数軸上データのパワースペクトル $|S(j)|$ と励起信号スペクトル $|E(j)|$ とを用いて上記(22)式のエラー ε_m を求め、その全バンドの総和値 $\sum \varepsilon_m$ を求めることができる。このエラー総和値 $\sum \varepsilon_m$ を各ピッチ毎に求め、最小となるエラー総和値に対応するピッチを最適のピッチとして決定

$$NSR = \frac{\sum_{j=a_m}^{b_m} \{ |S(j)| - |A_m| |E(j)| \}^2}{\sum_{j=a_m}^{b_m} |S(j)|^2} \quad \dots (24)$$

【0095】と表せ、このNSR値が所定の閾値(例えば0.3)より大のとき(エラーが大きい)ときには、そのバンドでの $|A_m| |E(j)|$ による $|S(j)|$ の近似が良くない(上記励起信号 $|E(j)|$ が基底として不適当である)と判断でき、当該バンドをUV(Unvoiced、無声音)と判別する。これ以外のときは、近似がある程度良好に行われていると判断でき、そのバンドをV

帯域)の下限、上限の点をそれぞれ a_m 、 b_m とすると、この第 m バンドのエラー ε_m は、

【0087】
【数10】

【0089】
10 【数11】

するわけである。以上のようにして高精度ピッチサーチ部106で最適のファイン(例えば0.25きざみ)ピッチが求められ、この最適ピッチに対応する振幅 $|A_m|$ が決定される。

【0092】以上ピッチのファインサーチの説明においては、説明を簡略化するために、全バンドが有声音(Voiced)の場合を想定しているが、上述したようにMBEボコーダにおいては、同時刻の周波数軸上に無声音(Unvoiced)領域が存在するというモデルを採用していることから、上記各バンド毎に有声音/無声音の判別を行うことが必要とされる。

【0093】上記高精度ピッチサーチ部106からの最適ピッチ及び振幅 $|A_m|$ のデータは、有声音/無声音判別部107に送られ、上記各バンド毎に有声音/無声音の判別が行われる。この判別のために、NSR(ノイズtoシグナル比)を利用する。すなわち、第 m バンドのNSRは、

【0094】
【数12】

(Voiced、有声音)と判別する。

【0096】次に、振幅再評価部108には、直交変換部105からの周波数軸上データ、高精度ピッチサーチ部106からのファインピッチと評価された振幅 $|A_m|$ との各データ、及び上記有声音/無声音判別部107からのV/UV(有声音/無声音)判別データが供給されている。この振幅再評価部108では、有声音/無声

音判別部107において無声音(UV)と判別されたバンドに関して、再度振幅を求めている。このUVのバンドについての振幅 $|A_m|_{uv}$ は、

$$|A_m|_{uv} = \sqrt{\sum_{j=a_m}^{b_m} |S(j)|^2 / (b_m - a_m + 1)} \quad \cdots (25)$$

【0098】にて求められる。

【0099】この振幅再評価部108からのデータは、データ数変換(一種のサンプリングレート変換)部109に送られる。このデータ数変換部109は、上記ピッチに応じて周波数軸上での分割帯域数が異なり、データ数(特に振幅データの数)が異なることを考慮して、一定の個数にするためのものである。すなわち、例えば有効帯域を3400Hzまでとすると、この有効帯域が上記ピッチに応じて、8バンド〜63バンドに分割されることになり、これらの各バンド毎に得られる上記振幅 $|A_m|$ (UVバンドの振幅 $|A_m|_{uv}$ も含む)データの個数 $m_{MX}+1$ も8〜63と変化することになる。このためデータ数変換部109では、この可変個数 $m_{MX}+1$ の振幅データを一定個数 N_C (例えば44個)のデータに変換している。

【0100】ここで本実施例においては、周波数軸上の有効帯域1ブロック分の振幅データに対して、ブロック内の最後のデータからブロック内の最初のデータまでの値を補間するようなダミーデータを付加してデータ個数を N_F 個に拡大した後、帯域制限型の K_{OS} 倍(例えば8倍)のオーバーサンプリングを施すことにより K_{OS} 倍の個数の振幅データを求め、この K_{OS} 倍の個数 $((m_{MX}+1) \times K_{OS})$ の振幅データを直線補間してさらに多くの N_M 個(例えば2048個)に拡張し、この N_M 個のデータを間引いて上記一定個数 N_C (例えば44個)のデータに変換する。

【0101】このデータ数変換部109からのデータ(上記一定個数 N_C の振幅データ)がベクトル量子化部110に送られて、所定個数のデータ毎にまとめられてベクトルとされ、ベクトル量子化が施される。ベクトル量子化部110からの量子化出力データは、出力端子111を介して取り出される。また、上記高精度のピッチサーチ部106からの高精度(ファイン)ピッチデータは、ピッチ符号化部115で符号化され、出力端子112を介して取り出される。さらに、上記有声音/無声音判別部107からの有声音/無声音(V/UV)判別データは、出力端子113を介して取り出される。これらの各出力端子111〜113からのデータは、所定の伝送フォーマットの信号とされて伝送される。

【0102】なお、これらの各データは、上記Nサンプル(例えば256サンプル)のブロック内のデータに対して処理を施すことにより得られるものであるが、プロ

【0097】

【数13】

ックは時間軸上を上記Lサンプルのフレームを単位として前進することから、伝送するデータは上記フレーム単位で得られる。すなわち、上記フレーム周期でピッチデータ、V/UV判別データ、振幅データが更新されることになる。

【0103】次に、伝送されて得られた上記各データに基づき音声信号を合成するための合成側(デコード側)の概略構成について、図14を参照しながら説明する。この図14において、入力端子121には上記ベクトル量子化された振幅データが、入力端子122には上記符号化されたピッチデータが、また入力端子123には上記V/UV判別データがそれぞれ供給される。入力端子121からの量子化振幅データは、逆ベクトル量子化部124に送られて逆量子化され、データ数逆変換部125に送られて逆変換され、得られた振幅データが有声音合成部126及び無声音合成部127に送られる。入力端子122からの符号化ピッチデータは、ピッチ復号化部128で復号化され、データ数逆変換部125、有声音合成部126及び無声音合成部127に送られる。また入力端子123からのV/UV判別データは、有声音合成部126及び無声音合成部127に送られる。

【0104】有声音合成部126では例えば余弦(cosine)波合成により時間軸上の有声音波形を合成し、無声音合成部127では例えばホワイトノイズをバンドパスフィルタでフィルタリングして時間軸上の無声音波形を合成し、これらの各有声音合成波形と無声音合成波形とを加算部129で加算合成して、出力端子130より取り出すようにしている。この場合、上記振幅データ、ピッチデータ及びV/UV判別データは、上記分析時の1フレーム(Lサンプル、例えば160サンプル)毎に更新されて与えられるが、フレーム間の連続性を高める(円滑化する)ために、上記振幅データやピッチデータの各値を1フレーム中の例えば中心位置における各データ値とし、次のフレームの中心位置までの間(合成時の1フレーム)の各データ値を補間により求める。すなわち、合成時の1フレーム(例えば上記分析フレームの中心から次の分析フレームの中心まで)において、先端サンプル点での各データ値と終端(次の合成フレームの先端)サンプル点での各データ値とが与えられ、これらのサンプル点間の各データ値を補間により求めるようにしている。

【0105】以下、有声音合成部126における合成処

理を詳細に説明する。上記V（有声音）と判別された第mバンド（第m高調波の帯域）における時間軸上の上記1合成フレーム（Lサンプル、例えば160サンプル）

$$V_m(n) = A_m(n) \cos(\theta_m(n)) \quad 0 \leq n < L \quad \dots (26)$$

と表すことができる。全バンドの内のV（有声音）と判別された全てのバンドの有声音を加算（ $\sum V_m(n)$ ）して最終的な有声音V(n)を合成する。

【0106】この(26)式中の $A_m(n)$ は、上記合成フレームの先端から終端までの間で補間された第m高調波の振幅である。最も簡単には、フレーム単位で更新され

$$A_m(n) = (L-n)A_{0m}/L + nA_{Lm}/L \quad \dots (27)$$

の式により $A_m(n)$ を計算すればよい。

$$\theta_m(0) = m\omega_{01}n + n^2 m(\omega_{L1} - \omega_{01})/2L + \phi_{0m} + \Delta\omega n \quad \dots (28)$$

により求めることができる。この(28)式中で、 ϕ_{0m} は上記合成フレームの先端（ $n=0$ ）での第m高調波の位相（フレーム初期位相）を示し、 ω_{01} は合成フレーム先端（ $n=0$ ）での基本角周波数、 ω_{L1} は該合成フレームの終端（ $n=L$ ：次の合成フレーム先端）での基本角周波数をそれぞれ示している。上記(28)式中の $\Delta\omega$ は、 $n=L$ における位相 ϕ_{Lm} が $\theta_m(L)$ に等しくなるような最小の $\Delta\omega$ を設定する。

【0108】以下、任意の第mバンドにおいて、それぞれ $n=0$ 、 $n=L$ のときのV/UV判別結果に応じた上記振幅 $A_m(n)$ 、位相 $\theta_m(n)$ の求め方を説明する。第mバンドが、 $n=0$ 、 $n=L$ のいずれもV（有声音）とされる場合に、振幅 $A_m(n)$ は、上述した(27)式により、伝送された振幅値 A_{0m} 、 A_{Lm} を直線補間して振幅 $A_m(n)$ を算出すればよい。位相 $\theta_m(n)$ は、 $n=0$ で $\theta_m(0) = \phi_{Lm} - m(\omega_{01} + \omega_{L1})L/2$

とし、かつ $\Delta\omega=0$ とする。

【0111】上記 $n=0$ 、 $n=L$ のいずれもV（有声音）とされる場合に、 $\theta_m(L)$ が ϕ_{Lm} となるように $\Delta\omega$

$$\begin{aligned} \theta_m(L) &= m\omega_{01}L + L^2 m(\omega_{L1} - \omega_{01})/2L + \phi_{0m} + \Delta\omega L \\ &= m(\omega_{01} + \omega_{L1})L/2 + \phi_{0m} + \Delta\omega L \\ &= \phi_{Lm} \end{aligned}$$

となり、これを整理すると、 $\Delta\omega$ は、

$$\Delta\omega = (\text{mod}2\pi((\phi_{Lm} - \phi_{0m}) - mL(\omega_{01} + \omega_{L1})/2)/L) \quad \dots (30)$$

となる。この(30)式で $\text{mod}2\pi(x)$ とは、 x の主値を $-\pi \sim +\pi$ の間の値で返す関数である。例えば、 $x=1.3\pi$ のとき $\text{mod}2\pi(x) = -0.7\pi$ 、 $x=2.3\pi$ のとき $\text{mod}2\pi(x) = 0.3\pi$ 、 $x=-1.3\pi$ のとき $\text{mod}2\pi(x) = 0.7\pi$ 、等である。

【0112】ここで、図15のAは、音声信号のスペクトルの一例を示しており、バンド番号（ハーモニクスナンバー）mが8、9、10の各バンドがUV（無声音）とされ、他のバンドはV（有声音）とされている。このV（有声音）のバンドの時間軸信号が上記有声音合成部126により合成され、UV（無声音）のバンドの時間軸

分の有声音を $V_m(n)$ とすると、この合成フレーム内の時間インデックス（サンプル番号） n を用いて、

る振幅データの第m高調波の値を直線補間すればよい。すなわち、上記合成フレームの先端（ $n=0$ ）での第m高調波の振幅値を A_{0m} 、該合成フレームの終端（ $n=L$ ：次の合成フレームの先端）での第m高調波の振幅値を A_{Lm} とすると、

【0107】次に、上記(26)式中の位相 $\theta_m(n)$ は、

$\theta_m(0) = \phi_{0m}$ から $n=L$ で $\theta_m(L)$ が ϕ_{Lm} となるように $\Delta\omega$ を設定する。

【0109】次に、 $n=0$ のときV（有声音）で、 $n=L$ のときUV（無声音）とされる場合に、振幅 $A_m(n)$ は、 $A_m(0)$ の伝送振幅値 A_{0m} から $A_m(L)$ で0となるように直線補間する。 $n=L$ での伝送振幅値 A_{Lm} は無声音の振幅値であり、後述する無声音合成の際に用いられる。位相 $\theta_m(n)$ は、 $\theta_m(0) = \phi_{0m}$ とし、かつ $\Delta\omega=0$ とする。

【0110】さらに、 $n=0$ のときUV（無声音）で、 $n=L$ のときV（有声音）とされる場合には、振幅 $A_m(n)$ は、 $n=0$ での振幅 $A_m(0)$ を0とし、 $n=L$ で伝送された振幅値 A_{Lm} となるように直線補間する。位相 $\theta_m(n)$ については、 $n=0$ での位相 $\theta_m(0)$ として、フレーム終端での位相値 ϕ_{Lm} を用いて、

を設定する手法について説明する。上記(24)式で、 $n=L$ と置くことにより、

信号が無声音合成部127で合成されるわけである。

【0113】以下、無声音合成部127における無声音合成処理を説明する。ホワイトノイズ発生部131からの時間軸上のホワイトノイズ信号波形を、所定の長さ（例えば256サンプル）で適当な窓関数（例えばハミング窓）により窓かけをし、STFT処理部132によりSTFT（ショートタームフーリエ変換）処理を施すことにより、図15のBに示すようなホワイトノイズの周波数軸上のパワースペクトルを得る。このSTFT処理部132からのパワースペクトルをバンド振幅処理部133に送り、図15のCに示すように、上記UV（無

声音)とされたバンド(例えば $m=8, 9, 10$)について上記振幅 $|A_m|_{uv}$ を乗算し、他の V (有声音)とされたバンドの振幅を0にする。このバンド振幅処理部133には上記振幅データ、ピッチデータ、 V/U V判別データが供給されている。バンド振幅処理部133からの出力は、ISTFT処理部134に送られ、位相は元のホワイトノイズの位相を用いて逆STFT処理を施すことにより時間軸上の信号に変換する。ISTFT処理部134からの出力は、オーバーラップ加算部135に送られ、時間軸上で適当な(元の連続的なノイズ波形を復元できるように)重み付けをしながらオーバーラップ及び加算を繰り返し、連続的な時間軸波形を合成する。オーバーラップ加算部135からの出力信号が上記加算部129に送られる。

【0114】このように、各合成部126、127において合成されて時間軸上に戻された有声音部及び無声音部の各信号は、加算部129により適当な固定の混合比で加算して、出力端子130より再生された音声信号を取り出す。

【0115】なお、上記図5の音声分析側(エンコード側)の構成や図14の音声合成側(デコード側)の構成については、各部をハードウェア的に記載しているが、いわゆるDSP(デジタル信号プロセッサ)等を用いてソフトウェアプログラムにより実現することも可能である。

【0116】また、本発明に係る有声音判別方法は、例えば、自動車電話の送信側で環境雑音(背景雑音等)を落としたいというとき、背景雑音を検出する手段としても用いられる。すなわち、雑音に乱された低品質の音声処理し、雑音の影響を取り除き、聞きやすい音にするようないわゆるスピーチエンハンスメントでの雑音検出にも適用される。

【0117】

【発明の効果】本発明に係る有声音判別方法は、信号の1ブロックをさらに分割した複数のサブブロック毎に求めた信号の統計的な性質の時間軸上での偏りに応じて有声音を雑音又は無声音かと区別することにより、確実に判別できる。そして、MBE等のボコーダに適用する場合には、音声のサブブロックに有声音入力がないとき、すなわち雑音又は無声音の入力があるとき、強制的に入力音声信号の全帯域を無声音として、間違ったピッチを検出することがないようにし、合成側での異音の発生を抑えることができる。

【0118】また、サブブロック毎の実効値(短時間rms値)の標準偏差及び平均値に基づいて短時間rms値の分布を調べることにより、少ない演算量で正確な有声音区間判別が行える。

【図面の簡単な説明】

【図1】本発明に係る有声音判別方法の第1の実施例を

説明するための有声音判別装置の概略構成を示す機能ブロック図である。

【図2】信号の統計的性質を説明するための波形図である。

【図3】第1の実施例を説明するための有声音判別装置の要部の構成を示す機能ブロック図である。

【図4】第1の実施例を説明するための有声音判別装置の要部の構成を示す機能ブロック図である。

【図5】本発明に係る有声音判別方法の第2の実施例を説明するための有声音判別装置の概略構成を示す機能ブロック図である。

【図6】本発明に係る有声音判別方法の第3の実施例を説明するための有声音判別装置の要部の概略構成を示す機能ブロック図である。

【図7】本発明に係る有声音判別方法の第4の実施例を説明するための有声音判別装置の概略構成を示す機能ブロック図である。

【図8】信号の統計的性質としての短時間rms値の分布を説明するための波形図である。

【図9】本発明に係る有声音判別方法が適用可能な装置の具体例としての音声信号の合成分析符号化装置の分析側(エンコード側)の概略構成を示す機能ブロック図である。

【図10】窓かけ処理を説明するための図である。

【図11】窓かけ処理と窓関数との関係を説明するための図である。

【図12】直交変換(FFT)処理対象としての時間軸データを示す図である。

【図13】周波数軸上のスペクトルデータ、スペクトル包絡線(エンベロープ)及び励起信号のパワースペクトルを示す図である。

【図14】本発明に係る有声音判別方法が適用可能な装置の具体例としての音声信号の合成分析符号化装置の合成側(デコード側)の概略構成を示す機能ブロック図である。

【図15】音声信号を合成する際の無声音合成を説明するための図である。

【符号の説明】

12・・・窓かけ処理部

13・・・サブブロック分割部

14・・・統計的性質検出部

15・・・標準偏差又は実効値情報検出部

16・・・ピーク値情報検出部

17・・・標準偏差又は実効値偏在検出部

18・・・判断部

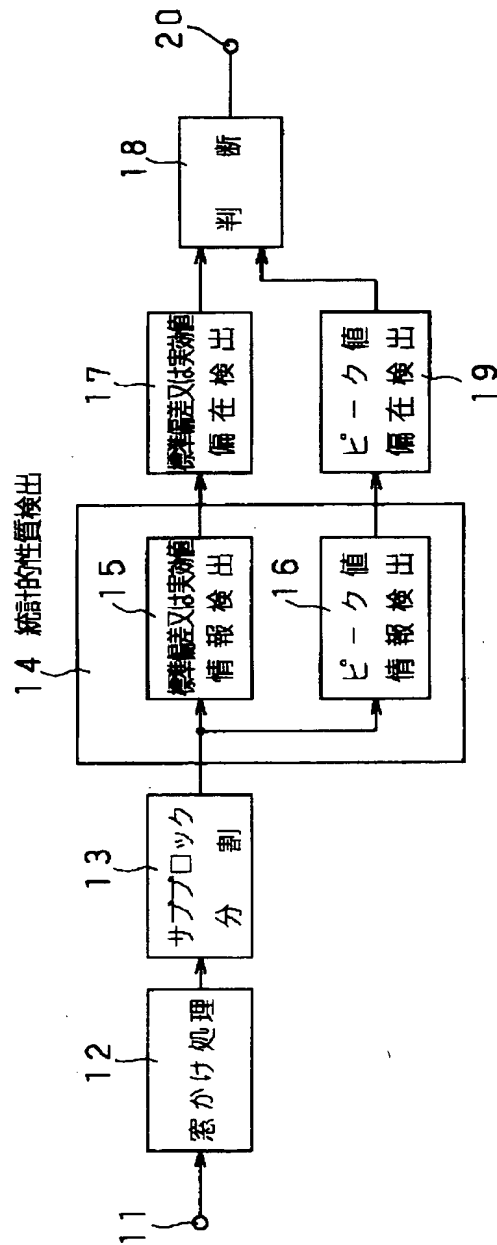
19・・・ピーク値偏在検出部

61・・・サブブロック毎の実効値算出部

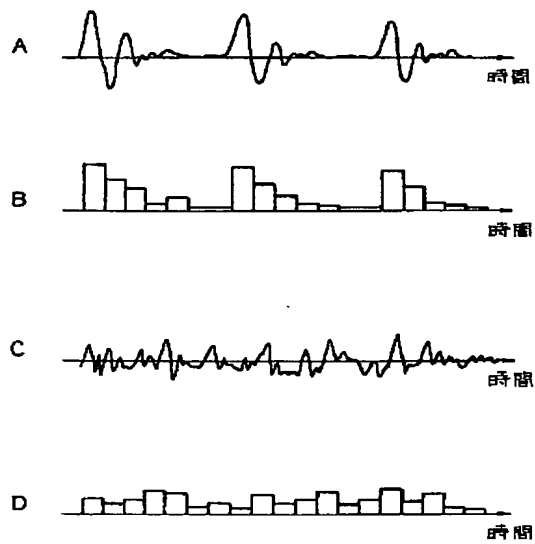
62・・・実効値の平均と標準偏差算出部

63・・・正規化された標準偏差算出部

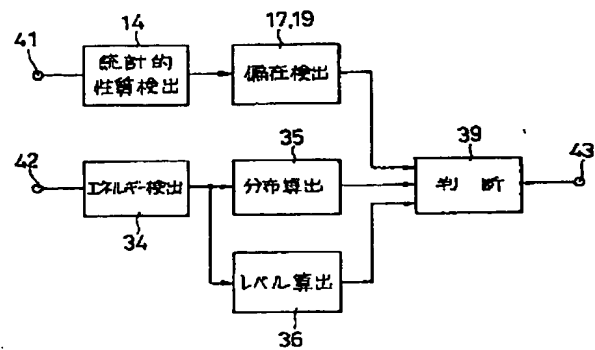
【図1】



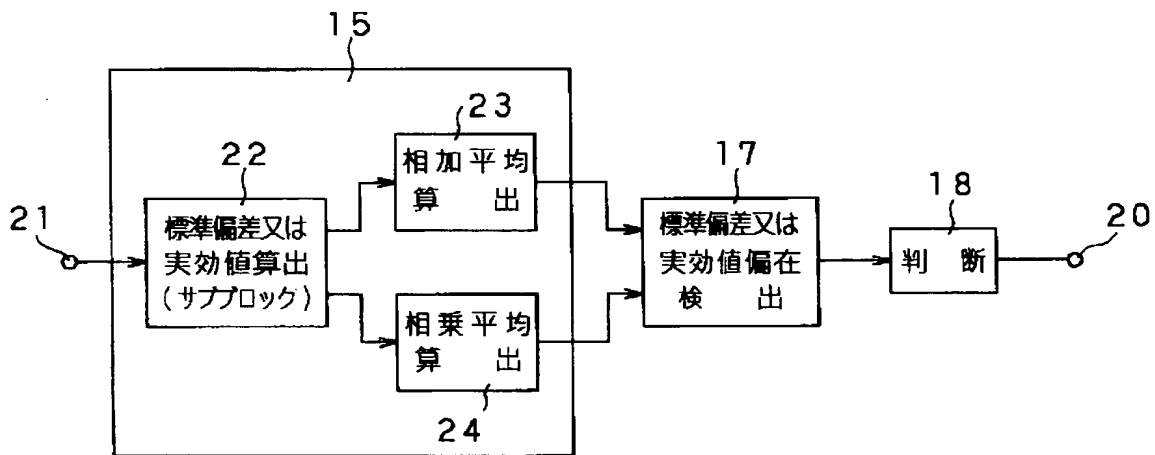
【図2】



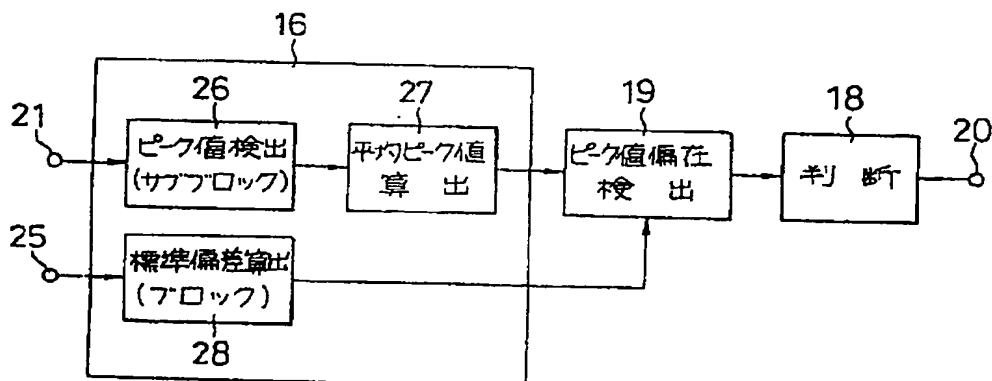
【図6】



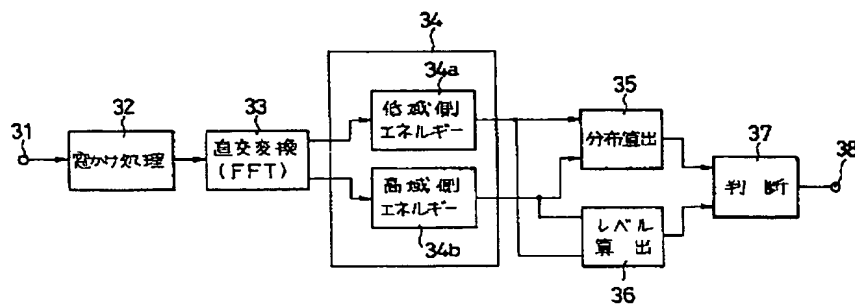
【図3】



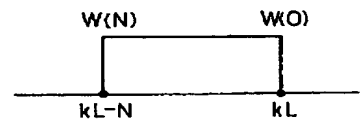
【図4】



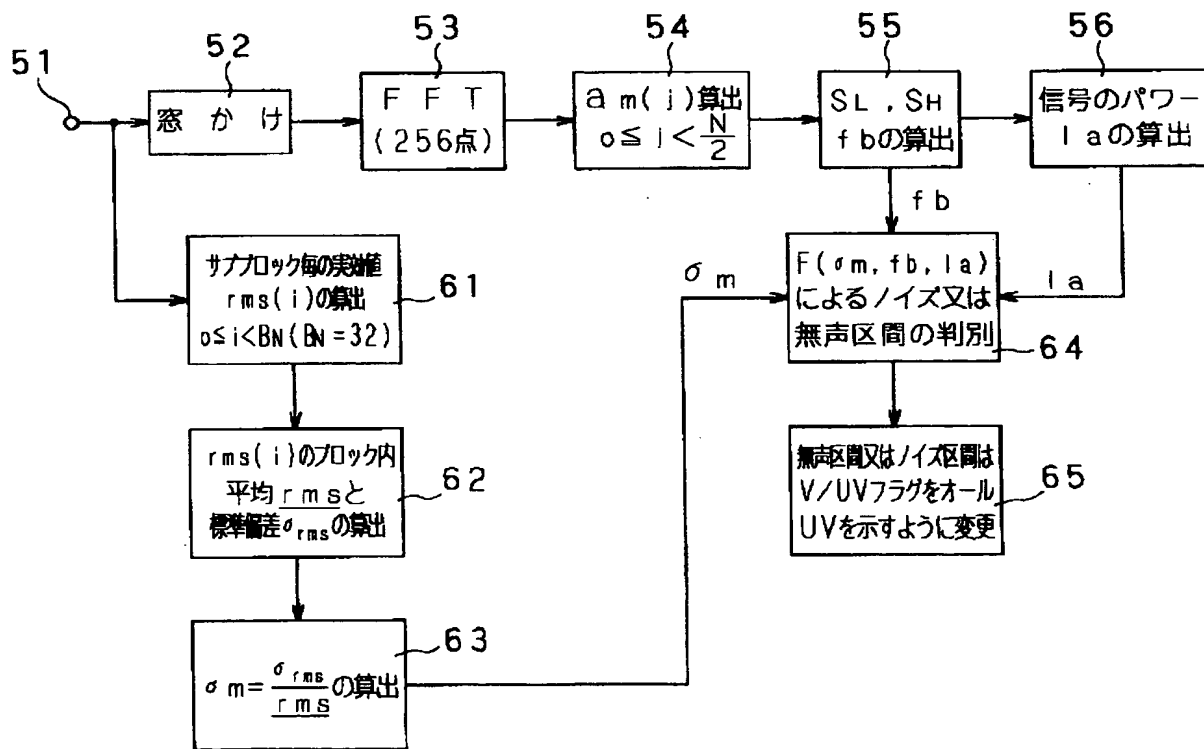
【図5】



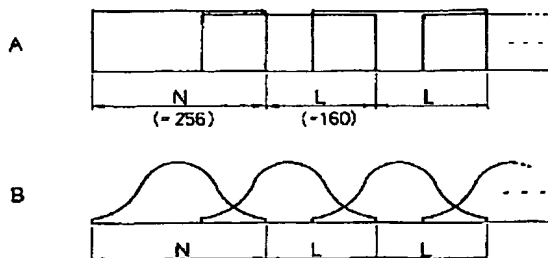
【図11】



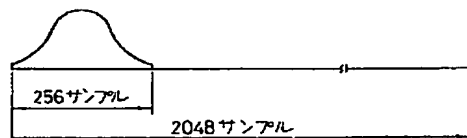
【図7】



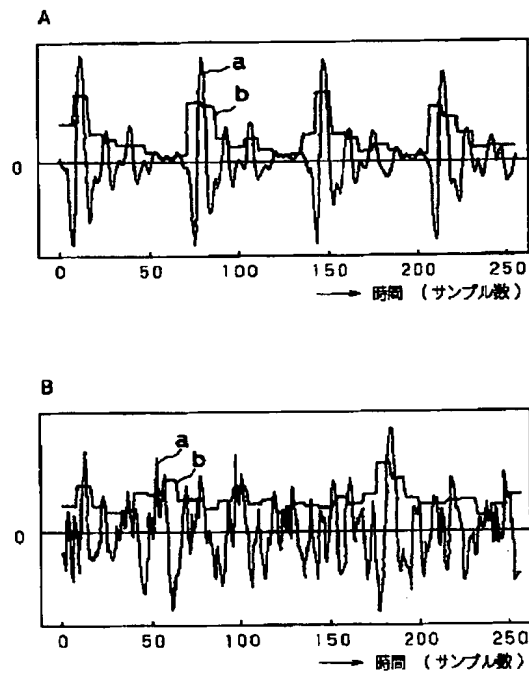
【図10】



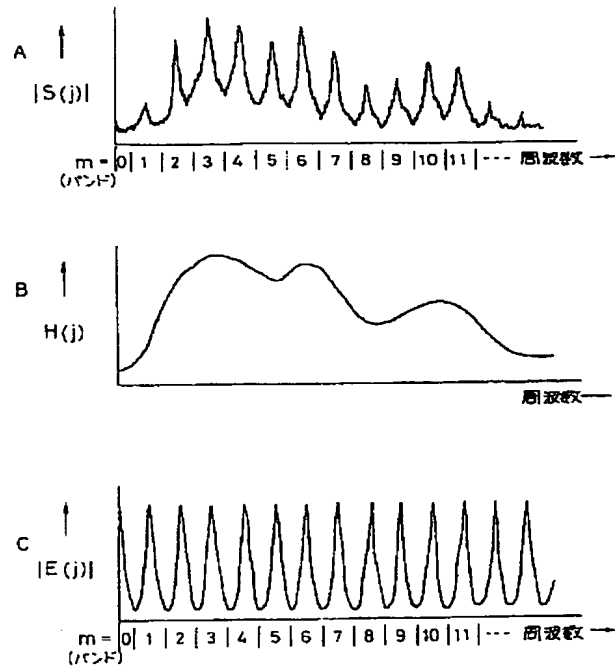
【図12】



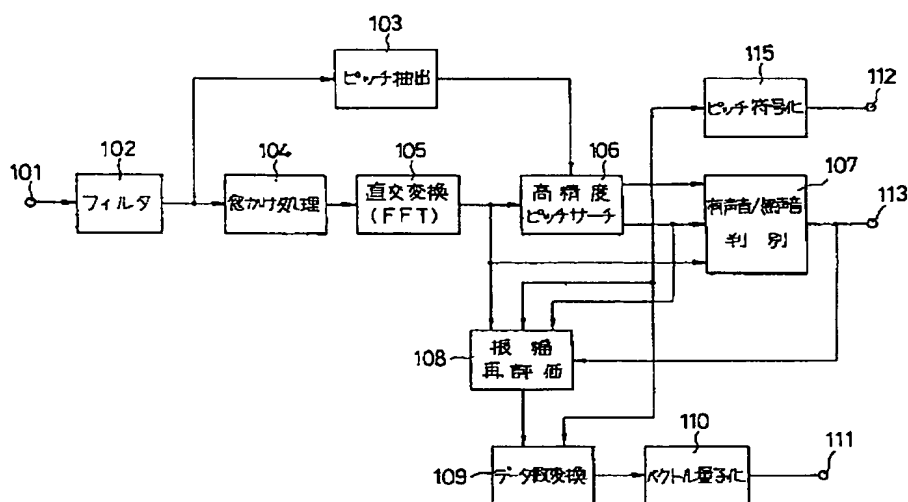
【図8】



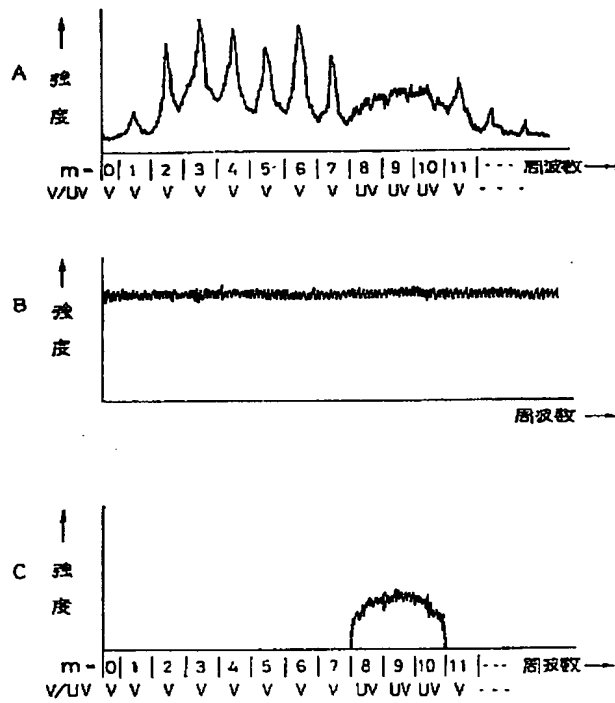
【図15】



【図9】



【図13】



【図14】

